Basira - Document Intelligence Platform

An automated workflow for ingesting, processing, and validating documents using Google Cloud services

Services Used

Cloud Storage

Used as a landing zone and for validation buckets.

Cloud Functions

Facilitates pipeline triggering and event-driven processing.

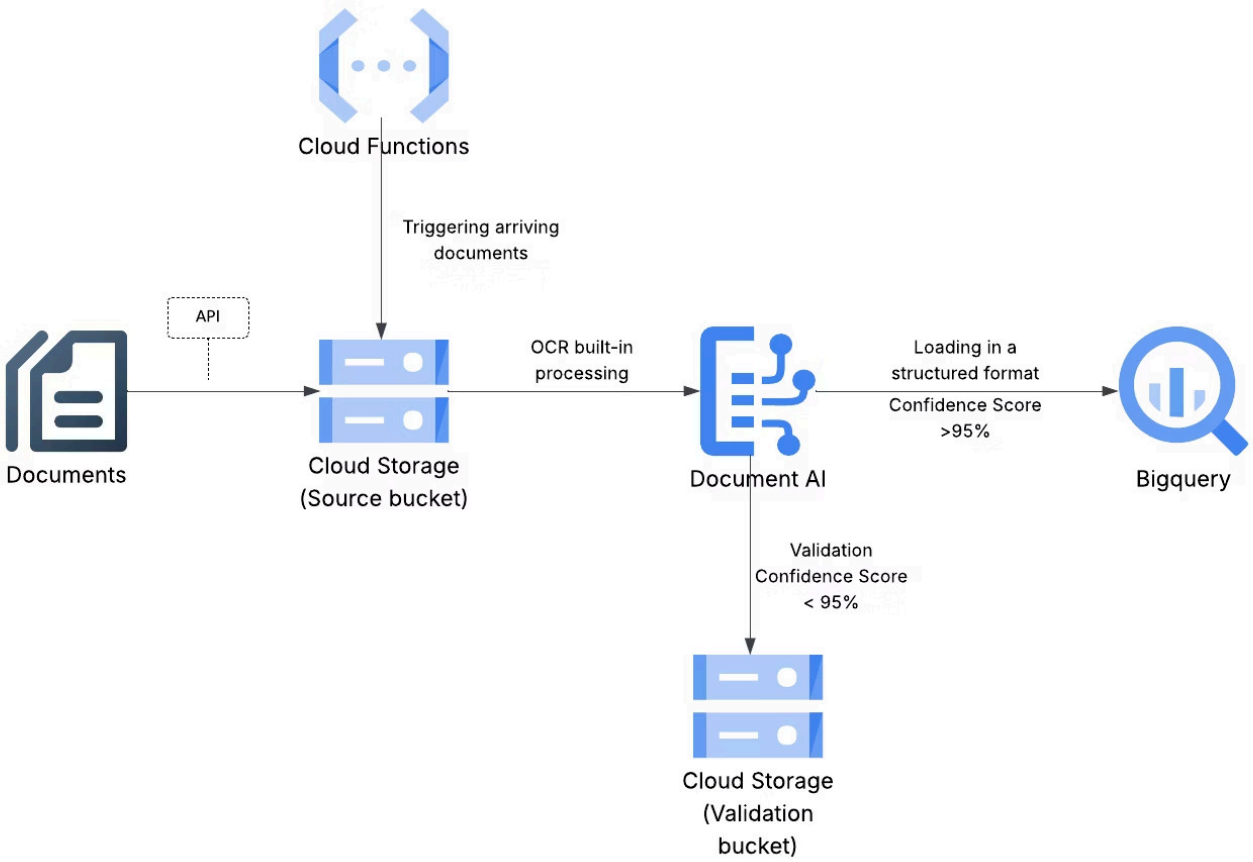
Document AI

Powers OCR capabilities and data extraction.

BigQuery

For structured storage and machine learning analytics.

Architecture Workflow



Complete end-to-end workflow showing document ingestion, processing, validation, and storage paths in the Basira Document Intelligence Platform.

The Journey Begins: Document Ingestion

External Source Ingestion

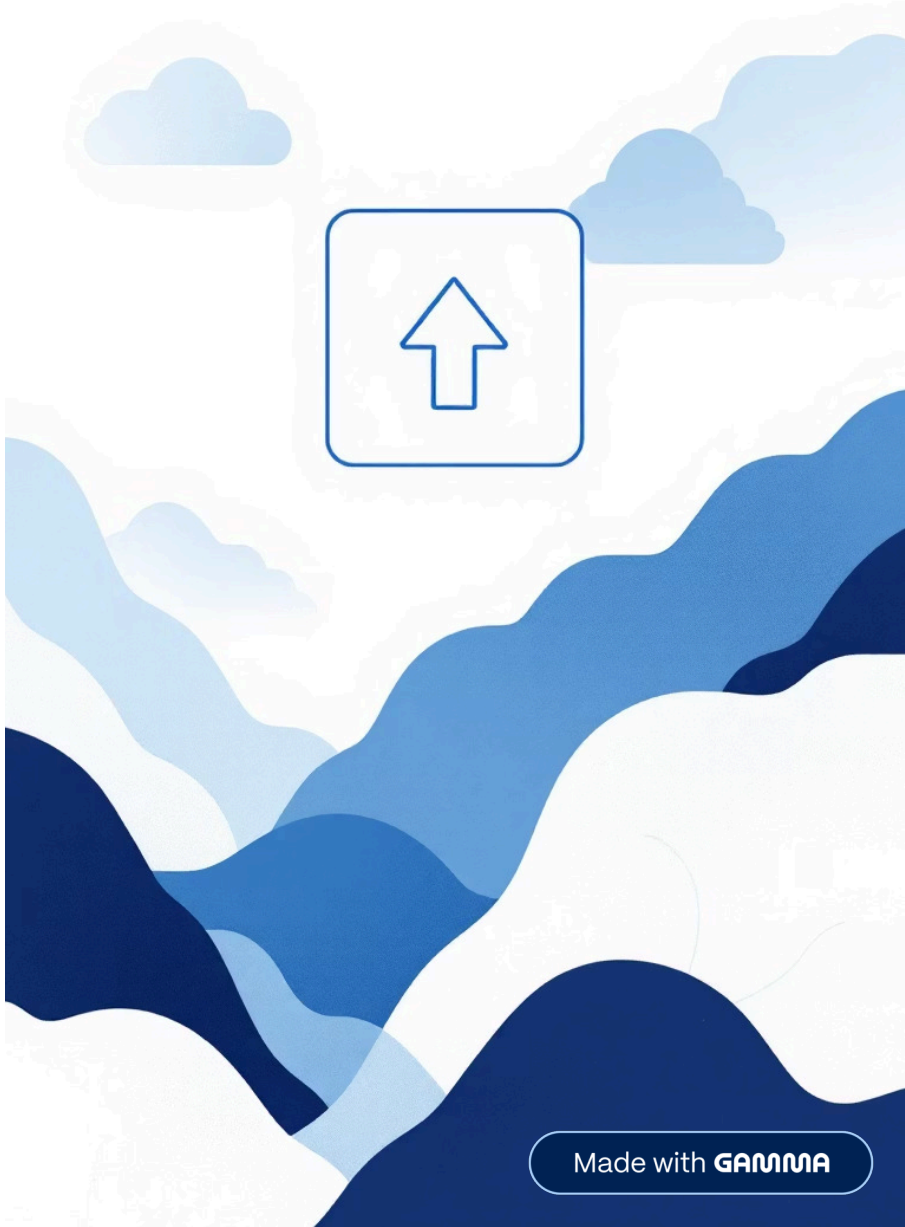
Documents from various external sources are uploaded via API directly into a designated Cloud Storage "Source Bucket" as the initial entry point.

Automated Processing Trigger

Upon arrival in the Source Bucket, an event-driven Cloud Function automatically triggers, initiating further processing by publishing messages to Pub/Sub.

Ingestion Challenges

Despite robust API design, potential issues like API reliability, network disruptions, corrupted files, or duplicate uploads must be addressed for seamless ingestion.



Process Triggering: Event-Driven Automation

Our document processing pipeline is initiated automatically through an event-driven mechanism, ensuring immediate action upon new document ingestion.

Source Bucket Monitoring

A dedicated Cloud Function continuously monitors the "Source Bucket" where new documents are initially uploaded.

Automatic Trigger Activation

Upon detecting a new document in the Source Bucket, the Cloud Function is automatically triggered, ensuring immediate response.

Processing Pipeline Initiation

The triggered Cloud Function then initiates the subsequent stages of the processing pipeline, such as publishing messages to Pub/Sub for further processing.

Serverless Trade-offs: Efficiency vs. Latency

While serverless functions offer significant benefits like scalability and cost-effectiveness, it's crucial to manage potential cold starts, which can introduce latency, and to design within execution timeout limits for complex processing tasks.

Data Extraction (OCR)

Document AI Integration

The Cloud Function forwards documents to Google Cloud's **Document AI** service for advanced optical character recognition (OCR) and intelligent data extraction.

Text & Data Extraction

Document AI processes various file types, extracting raw text and identifying key data fields, transforming unstructured content into a machine-readable format.

Structured Output

Extracted data is provided in a structured format, enabling direct integration with downstream systems like databases or analytics platforms.

OCR Performance & Confidence

Accuracy and performance vary based on document type. Text-based documents generally offer higher precision than scanned images. Specialized processors for structured documents like invoices and bank statements yield significantly higher confidence scores.

Confidence Score & Validation

High Confidence (>95%)

Documents with extraction confidence above 95% are automatically processed and routed directly to **BigQuery** for storage and analysis.

Low Confidence (<95%)

Documents with lower confidence scores are flagged and directed to a validation bucket, awaiting human-in-the-loop review for accuracy correction.

Threshold Selection

The 95% threshold is a configurable parameter. Optimizing this balance is crucial for system efficiency and data quality, reflecting the acceptable risk for automated processing.

Speed vs. Accuracy

A higher threshold ensures greater accuracy but may increase human review volume. A lower threshold boosts processing speed but risks lower data quality without manual oversight.

Operational Overhead

The trade-off directly impacts operational costs; extensive human review means higher labor expenses, while fully automated processes require robust error handling and monitoring.

This tiered approach ensures data integrity while optimizing the balance between automated efficiency and human oversight, delivering reliable, validated records to **BigQuery** for robust analysis.

Key Benefits & Trade-offs

Benefits

Automated Processing

Streamlined workflows reduce manual effort and accelerate data processing.

Scalable Architecture

Flexibly handles varying data volumes, ensuring consistent performance.

Intelligent Validation

AI-powered validation enhances accuracy and reduces errors.

Continuous Improvement

Human-in-the-Loop feedback refines models and improves outcomes over time.

Trade-offs

API Dependency

Reliance on external APIs introduces potential points of failure and vendor lock-in.

Cold Start Delays

Initial processing may experience delays due to resource spin-up.

Threshold Tuning

Optimizing confidence thresholds for accuracy requires careful calibration.

Manual Review Overhead

Some level of human review remains necessary, adding to operational costs.